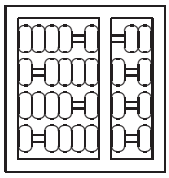

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Um algoritmo para identificação de correlações múltiplas de polimorfismos

André Atanasio M. Almeida *Miguel Galves* *Zanoni Dias*

Technical Report - IC-06-14 - Relatório Técnico

September - 2006 - Setembro

Um algoritmo para identificação de correlações múltiplas de polimorfismos

André Atanasio M. Almeida * Miguel Galves † Zanoni Dias ‡

Resumo

Polimorfimo de Base Única (SNP) é uma mutação que afeta apenas uma posição do genoma de um organismo. Correlação de polimorfismos (LD) é uma associação não aleatória de SNPs, podendo ser empregada como marcador. Correlação múltipla de polimorfismos agrupa três ou mais polimorfismos, permitindo que quaisquer dois polimorfismos sejam utilizados como marcadores. Neste trabalho, estudamos duas definições de LDs (LD completo e LD útil) e apresentamos uma definição para LDs múltiplos fundamentada em teoria dos grafos, assim como uma heurística gulosa, baseada no grau dos vértices, para sua identificação. São exibidos resultados promissores obtidos em testes com dados do genoma da cana-de-açúcar e do genoma humano.

1 Introdução

Dizemos que há um polimorfismo em uma seqüência genética quando existe uma ou mais formas genéticas (alelos) distintas em indivíduos da mesma espécie. Para que um alelo seja considerado um polimorfismo, ele deve aparecer em, pelo menos, 1% da população analisada. Caso contrário, considera-se que o alelo é uma mutação pontual.

Polimorfismo de Base Única (SNP – *Single Nucleotide Polymorphism*, em inglês) é um polimorfismo que ocorre em apenas uma posição do genoma [4]. Não são considerados SNPs as inserções ou remoções simples de bases em uma seqüência genômica.

Os SNPs podem ser polimorfismos bi, tri ou tetra alélicos, ou seja, possuirem duas, três ou quatro formas distintas. Porém, os dois últimos tipos são raros. As variações mais freqüentes são substituições entre bases nitrogenadas de mesma caraterística estrutural (um A por um G, um G por um A, um C por um T ou ainda um T por um C). Tais substituições chamamos de transições. As outras substituições chamamos de transversões.

Podemos classificar um SNP como sinônimo ou não. Dizemos que o SNP é sinônimo se o aminoácido codificado pelo *codon* contendo o SNP é o mesmo que aquele codificado pelo *codon* sem SNP e é não sinônimo, caso contrário. Um SNP não sinônimo pode modificar a estrutura e a função da proteína codificada. Um dos maiores interesses da pesquisa sobre o genoma humano é determinar se um SNP não sinônimo, chamado de nsSNP, afeta a função

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
†Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
‡Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

da proteína e conseqüentemente tem impacto sobre a saúde do indivíduo. Aproximadamente metade das causas genéticas de doenças originam-se da substituição de aminoácidos [5].

O estudo de SNPs começa com a coleta de seqüências de uma região comum obtidas de diversos indivíduos de uma população. Diversas técnicas podem ser empregadas para a obtenção das seqüencias, dentre as quais podemos citar PCR [16] e EST [1]. Em seguida as seqüências são submetidas a um alinhamento múltiplo [17, Seção 3.4]. Ao resultado do alinhamento damos o nome de *contig*.

1.1 Correlação de polimorfismos

Correlação de polimorfismos ou, ainda, desequilíbrio de ligação (LD – do inglês *Linkage Disequilibrium*) é uma associação não aleatória de alelos.

Um conjunto de SNPs estatisticamente associados é chamado de haplótipo. Quando dois ou mais alelos específicos, em *loci* distintos, em um mesmo cromossomo são sempre encontrados em conjunto, então os *loci* estão em desequilíbrio de ligação [15, pg. 177]. Quando isto acontece, a identificação de um um SNP em um locus fornece informações sobre SNPs em outros loci.

Análises de LD são mais efetivas em populações isoladas, que em geral possuem menor heterogeneidade alélica, ou em análises de doenças causadas por mutações mais antigas e bastante comuns na população humana. Como exemplo de doença que se aplica ao primeiro caso podemos citar a displasia distrófica em finlandeses e como exemplo do segundo caso podemos citar a fibrose cística e a doença de Huntington em populações européias [3].

Correlacionar um fenótipo observado com um genótipo é um dos objetivos fundamentais da genética. Obter a seqüência genética relacionada a uma doença é essencial para se produzir terapias e tratamentos adequados. Métodos gerais para descobrir os genes responsáveis por doenças, que possuem características mendelianas simples, só foram determinadas no início da década de 80, quando a análise de ligação de genes foi proposta pela primeira vez.

Genes que controlam características mendelianas podem ser identificados e isolados com base em informações acerca de características hereditárias, utilizando a técnica de clonagem posicional [3].

As técnicas tradicionais, tais como a clonagem posicional, são capazes de delimitar regiões cromossomais de 1 a 2 cM (um a dois *centimorgans*). Para um mapeamento mais fino de genes, é recomendável utilizar técnicas baseadas em desequilíbrio de ligação, que permitem marcar regiões menores e mais próximas dos genes de interesse. Tais técnicas permitem a delimitação de regiões de $0,1$ cM, o que é equivalente a aproximadamente 100 kbp (cem mil pares de base) quando falamos de genoma humano.

Muito do conhecimento e entendimento de como LDs são formados na natureza veio do estudo feito em espécies de *Drosophila* [2] e, em particular, da *Drosophila melanogaster*, onde foram feitos estudos mais detalhados de LDs.

1.2 Métricas para LD

Diversas medidas foram criadas para se quantificar LD [7]. A mais antiga das medidas propostas para LD é chamada D. Esta medida quantifica um LD como sendo a diferença

entre freqüência observada entre haplótipos de dois *loci* e a freqüência que seria esperada se os alelos fossem aleatórios. Considerando os alelos A, a, B e b temos:

$$D = P_{AB} - P_A P_B$$

onde P_A e P_B são as probabilidades de aparição dos alelos separadamente e P_{AB} é a probabilidade dos dois alelos aparecerem juntos. Podemos afirmar que: $P_A = P_{AB} + P_{Ab}$, $P_a = P_{aB} + P_{ab}$, $P_B = P_{AB} + P_{aB}$, $P_b = P_{Ab} + P_{ab}$ e $P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$. Observe que:

$$P_A P_B = P_{AB}(P_{AB} + P_{Ab} + P_{aB}) + P_{Ab}P_{aB}$$

então:

$$\begin{aligned} P_{AB} - P_A P_B &= P_{AB} - P_{AB}(P_{AB} + P_{Ab} + P_{aB}) - P_{Ab}P_{aB} \\ &= P_{AB}(1 - P_{AB} - P_{Ab} - P_{aB}) - P_{Ab}P_{aB} \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB} \end{aligned}$$

O valor numérico de D tem pouco uso na comparação de LDs. Isto se deve ao fato de ser dependente da freqüência de alelos, o que dificulta a comparação e a avaliação dos resultados. Sendo assim, várias outras medidas baseadas em D, com escala variando de 0 a 1, foram propostas.

Uma das medidas baseadas em D é D', proposta por Lewontin [13]. Tal medida é dada pela seguite fórmula:

$$D' = \frac{|D|}{D_{max}}$$

com:

$$D_{max} = \begin{cases} min(P_A P_b, P_a P_B) & \text{se } D > 0 \\ min(P_A P_B, P_a P_b) & \text{se } D < 0 \end{cases}$$

O denominador da fórmula corresponde ao maior valor que D pode assumir dadas as probabilidades de aparição de cada alelo. Quando $D' = 1$ dizemos que o LD é completo, ou seja, os dois SNPs não foram separados por recombinação. Valores intermediários de D' são de difícil interpretação.

Outra medida baseada em D é r^2, também denotada por Δ^2. Tal medida é apontada por Hill e Weir [8] como sendo a mais utilizada para análise de LDs. Sua fórmula é a seguinte:

$$r^2 = \frac{D^2}{P_A P_a P_B P_b}$$

Quando $r^2 = 1$ dizemos que o LD é perfeito. Tal situação acontece se e somente se os marcadores não foram separados por recombinação e têm a mesma freqüência alélica.

A medida r^2 tem sido muito utilizada para se definir o que são LDs úteis [12]. De fato, o aumento do número de amostras em estudos de associação tem custo alto, e aumentar o

número de amostras para compensar LDs fracos é praticamente inviável. LDs com $r^2 \geq 1/3$ são considerados como úteis em processos de mapeamentos.

Diversas outras métricas existem, porém como D' e r^2 são as mais utilizadas e conhecidas, decidimos por limitar nosso estudo a estas.

A definição de LD é bastante abrangente, permitindo sua aplicação a qualquer tipo de polimorfismo. No contexto deste trabalho, quando falarmos em LD, estaremos sempre falando de um LD relativo a SNPs.

Na Seção 2 apresentamos o conceito de LD múltiplo e o algoritmo que sugerimos para o seu cálculo. Na Seção 3 apresentamos os resultados obtidos com a aplicação do algoritmo, por nós proposto, em dados oriundos do Projeto `SUCEST` (cana-de-açúcar). Apresentamos o resultado da aplicação de tal algoritmo em dados do genoma humano na Seção 4. Finalmente, na Seção 5 expomos a conclusão.

2 LDs Múltiplos

Dado um *contig*, podemos construir um grafo [21] onde cada vértice representa um SNP. Se dois SNPs definem um LD então os vértices correspondentes são ligados por uma aresta. A Figura 1 nos apresenta um exemplo representando 16 SNPs.

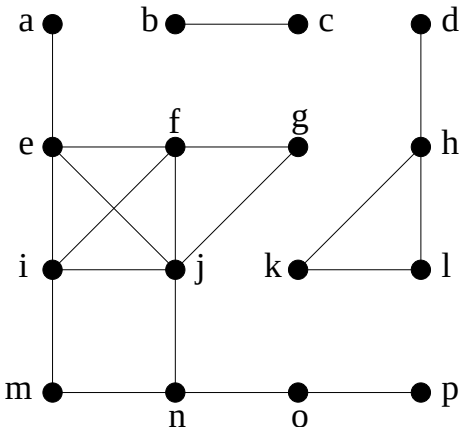

Figura 1: Um exemplo de grafo representando 16 SNPs.

Definimos tamanho de um grafo como o número de vértices no grafo. Dizemos que o vértice a é vizinho do vértice b caso exista uma aresta ligando a e b. Definimos vizinhança de um vértice como o conjunto de vértices que são vizinhos ao vértice. Definimos grau de um vértice como a cardinalidade de sua vizinhança, ou em outras palavras, como o número de vértices vizinhos. Definimos caminho como uma seqüência de vértices tal que para cada vértice na seqüência, exceto o último, exista uma aresta que o ligue com seu sucessor. Seja C um conjunto. Dizemos que o conjunto $S1 \subset C$ é maximal, em C, para a propriedade P se $S1$ atende P e não existe um conjunto $S2 \neq S1$ tal que $S1 \subset S2 \subset C$ e $S2$ atende P. Dizemos que $S1 \subset C$ é máximo, em C, para a propriedade P se $S1$ atente P e não existe um

conjunto $S2 \subset C$ tal que $S2$ atende P e a cardinalidade de $S2$ é maior que a cardinalidade de $S1$.

Uma componente conexa de um grafo é um conjunto maximal de vértices tal que existe pelo menos um caminho que liga cada par de vértices. Por exemplo, no grafo da Figura 1 temos três componentes conexas, a saber: $\{b, c\}$, $\{d, h, k, l\}$ e $\{a, e, f, g, i, j, m, n, o, p\}$. Uma clique em um grafo é um conjunto de vértices tal que cada vértice é vizinho de todos os outros do conjunto. São exemplos de clique no grafo da Figura 1: $\{b, c\}$, $\{h, k, l\}$, $\{f, g, j\}$ e $\{e, f, i, j\}$. Observe que todas estas são cliques maximais e que $\{e, f, i, j\}$ é a clique máxima no grafo. Definimos o LD múltiplo de um SNP como a clique máxima, de tamanho maior ou igual a três, que contém o SNP, ou em outras palavras, o LD múltiplo do SNP x é o maior conjunto S de SNPs tal que $x \in S$ e existe um LD entre todo par de elementos de S.

Os LDs (simples) podem ser empregados como marcadores de regiões cromossômicas. Os LDs múltiplos permitem que qualquer par de seus SNPs sejam empregados como marcadores de regiões cromossômicas.

Na Seção 2.1 apresentamos como calcular os LDs múltiplos de um dado *contig*.

2.1 Cálculo dos LDs múltiplos

Dado um *contig*, construímos seu respectivo grafo. Calculamos as componentes conexas e, então, passamos a buscar pelos LDs múltiplos para cada SNP.

Objetivando encontrar o LD múltiplo para cada SNP, realizamos o seguinte procedimento. Para cada componente conexa, buscamos pela maior clique. Definimos esta como a maior clique para cada um de seus vértices. Em seguida, passamos a buscar pela maior clique dos demais vértices na componente conexa.

A busca por cliques máximas em grafos é um problema NP-Difícil [11], ou seja, é um problema para o qual não é conhecido um algoritmo eficiente em termos computacionais. Sendo assim, decidimos por adotar uma heurística gulosa baseada no grau dos vértices no procedimento de busca por clique máxima em componentes conexas. A busca é realizada da seguinte forma:

- Eliminamos os vértices de grau um. Esta etapa é repetida diversas vezes até que todos os vértices de grau um sejam eliminados, uma vez que a remoção de um vértice de grau um pode gerar um novo. Tal procedimento irá contribuir com a redução no número de possibilidades a verificar. Observe que tal operação não afeta a busca por cliques de tamanho maior ou igual a três (LDs múltiplos).

- Criamos uma lista de vértices ordenada pelos seus respectivos graus. Esta lista é utilizada, mais uma vez, na redução do número de possibilidades. Observe que, se buscamos por uma clique de tamanho n, um vértice de grau menor que $n - 1$ não pode estar presente.

- Seja d o maior grau entre os vértices. Começamos por buscar uma clique com $d + 1$ vértices durante t segundos. Procuramos por todos os vértices com grau maior ou igual a d (a lista ordenada por graus facilita a busca). Digamos que foram encontrados n vértices. Se $n \geq d+1$ então há possibilidade de encontrarmos a clique e assim fazemos

as combinações destes n vértices em grupos de $d+1$ elementos. Para cada grupo de $d+1$ SNPs, verificamos se é clique. Na primeira resposta positiva o procedimento é encerrado retornando a clique corrente. Caso esta busca por uma clique, com $d+1$ vértices, exceda t segundos o procedimento é encerrado sem nada retornar. Caso não seja encontrada a clique com $d+1$ vértices, passamos a buscar por uma clique com d vértices novamente durante t segundos e assim sucessivamente até encontrar uma clique. Em último caso, uma clique com dois vértices é encontrada. O limite de tempo nas buscas serve para limitar o processo. Observe que cada vez que o procedimento atinge o limite de tempo a busca torna-se mais flexível, cliques menores são buscadas.

Na Figura 2 apresentamos o resultado da aplicação do primeiro passo do algoritmo na maior componente conexa do exemplo que exibimos na Figura 1. Pelo segundo passo temos a seguinte lista de vértice:

$$j \quad f \quad i \quad e \quad g \quad m \quad n$$

com seus respectivos graus:

$$5 \quad 4 \quad 4 \quad 3 \quad 2 \quad 2 \quad 2\ .$$

No terceiro passo começamos a buscar por uma hipotética clique com 6 vértices, uma vez que o maior grau é 5. Como apenas um dos vértices tem grau maior ou igual a 5 então podemos afirmar que tal clique não existe. Passamos a buscar por uma clique de tamanho 5. Também não é possível encontrar. Há apenas três vértices com grau maior ou igual a 4. Passamos então a buscar uma clique de tamanho 4. Desta vez há possibilidade, pois temos quatro vértices (e, f, i e j) com grau maior ou igual a 3. Esta é realmente uma clique e assim encerramos nossa busca. Observe que neste caso tivemos apenas uma possibilidade a verificar, mas isto nem sempre ocorre. Por exemplo, caso estivéssemos procurando por uma clique de tamanho 3 teríamos 35 possibilidades a analisar, pois há sete vértice com grau maior ou igual a 2.

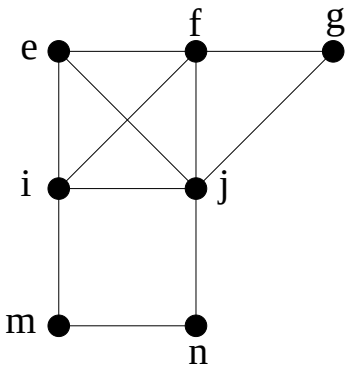

Figura 2: Apresenta o resultado da aplicação do primeiro passo do algoritmo de busca pela clique máxima na maior componente conexa do exemplo da Figura 1.

Uma vez encontrada a maior clique na componente conexa, um procedimento semelhante é usado para buscar pela maior clique de cada um dos vértices restantes. A diferença está no fato de termos um elemento fixo na clique e os testes, na busca por clique, limitam-se ao elemento e seus vizinhos. Como adotamos uma heurística, nosso algoritmo não garante que a maior clique de cada vértice foi encontrada.

Interpretamos que vértices com maior clique de tamanho dois são LDs simples e que vértices com maior clique de tamanho um são SNPs isolados.

3 LDs múltiplos no genoma da cana-de-açúcar

O Projeto SUCEST (*The Sugarcane EST Project*) [19, 20] foi um projeto brasileiro de seqüenciamento de ESTs de cana-de-açúcar realizado por uma rede de laboratórios de pesquisa financiados pela FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo).

Os ESTs de cana-de-açúcar com SNPs anotados foram extraídos da base de dados do projeto SUCEST. Inicialmente, um conjunto de 291689 ESTs foi produzido. Tal conjunto é composto por seqüências com um tamanho médio de $829,44 \pm 182,60$ bp com qualidade média de $23,15 \pm 15,71$. Posteriormente, as sequências genéticas foram agrupadas em *clusters* utilizando o pacote `cap3` [10]. Foram gerados 43141 *clusters* onde 16338 são *singlets* (*clusters* formados por um único EST).

Um estudo sobre SNPs foi realizado em tais *clusters* e foram catalogados SNPs em 8198 *contigs*. Realizamos, neste conjunto de dados, um estudo acerca de LDs múltiplos tal como descrito a seguir.

Inicialmente selecionamos apenas os SNPs bialélicos. Depois selecionamos os *contigs* onde havia, pelo menos, dois SNPs. Com isto, nos restaram 6178 *contigs*. Nestes dados aplicamos o procedimento descrito na Seção 2.1.

Na Seção 3.1 definimos as métricas e os limiares utilizados para definir LD. Resultados de testes com diferentes configurações do parâmetro t do algoritmo são mostrados na Seção 3.2. Finalmente, na Seção 3.3 apresentamos os resultados obtidos com a aplicação de nosso algoritmo em dados oriundos do Projeto `SUCEST`.

3.1 Definição de LD

Durante o procedimento de seqüenciamento de material genético é atribuído um valor de qualidade a cada base seqüenciada. Este número relaciona-se com a probabilidade daquela leitura estar correta. Podemos definir um limiar e classificar as bases como sendo de baixa qualidade, caso tenha valor inferior ao limiar, ou de alta qualidade, caso contrário.

Anteriormente dissemos que uma aresta é criada sempre que dois SNPs formam um LD. Mas quando considerar que dois SNPs formam um LD? Das diversas métricas existentes, consideramos apenas duas: D' e r^2. Em nosso estudo, fizemos uma análise utilizando quatro critérios distintos. São eles:

- Consideramos LD apenas quando $D' = 1$ e as bases de baixa qualidade foram desprezadas nos cálculos.

- Consideramos LD apenas quando $D' = 1$ e as bases de baixa qualidade foram utilizadas nos cáculos tal como as de alta qualidade.
- Consideramos LD apenas quando $r^2 \geq 1/3$ e as bases de baixa qualidade foram desprezadas nos cálculos.
- Consideramos LD apenas quando $r^2 \geq 1/3$ e as bases de baixa qualidade foram utilizadas nos cálculos tal como as de alta qualidade.

Observe que em cada um dos casos temos, possivelmente, grafos distintos.

3.2 Verificando limite para o tempo de busca por clique

A Tabela 1 nos mostra o tempo total e o tempo médio por *contig* gasto na busca por LDs múltiplos utilizando limite de tempo para busca por cliques máximas de 5 e 60 segundos. Mostra-nos também o quão mais lenta foi a busca com limite de 60 segundos. Em geral, usar a métrica r^2 é cerca de 10 vezes mais rápido do que usar a métrica D. Observe que, na comparação entre $t = 5$ e $t = 60$ segundos, o tempo, quando utilizamos r^2 como métrica, cresce aproximadamente 100% enquanto que quando utilizamos D' o tempo é 5 a 6 vezes o tempo inicial. Apresentamos os valores para cada situação: $r^2 \geq 1/3$ eliminando bases de baixa qualidade, $r^2 \geq 1/3$ considerando bases de baixa qualidade, $D' = 1$ eliminando bases de baixa qualidade e $D' = 1$ considerando bases de baixa qualidade. A seguir uma descrição detalhada de cada coluna da tabela:

- A coluna **Métrica** indica a situação a que se referem os valores. O símbolo * indica que consideramos as bases de baixa qualidade.
- A coluna **T5** indica o tempo total gasto em segundos para calcular os LDs múltiplos de todos os *contigs* com limite de tempo de 5 segundos para a busca por clique de determinado tamanho.
- A coluna **T5/c** apresenta o tempo médio gasto por *contig* para cálculo de LDs múltiplos quando limitamos o tempo de busca em 5 segundo.
- As colunas **T60** e **T60/c** são semelhantes a **T5** e **T5/c** só que a limitação do tempo de busca agora é 60 segundos.
- Na coluna **T60/T5** é exibido o valor da divisão de **T60** por **T5**.

Realizamos estes dois testes, com limite de $t = 5$ e $t = 60$ segundos, para avaliar se $t = 5$ segundos é suficiente para respostas satisfatórias. Para realizar a comparação dos resultados observamos parâmetros relativos às componentes conexas e outros relativos a vértices (SNPs).

Quanto a componentes conexas, foram três os parâmetros observados: tamanho da maior clique, número de cliques de tamanho 1 e número de cliques de tamanho 2. Não foram observadas diferenças em qualquer dos quatro casos abordados.

Métrica	**T5**	**T5/c**	**T60**	**T60/c**	**T60/T5**
$r^2 \geq 1/3$	$325,314$	$0,053$	$740,991$	$0,120$	$2,278$
$r^2 \geq 1/3$ *	$345,609$	$0,056$	$750,849$	$0,121$	$2,172$
$D' = 1$	$2110,830$	$0,342$	$10386,839$	$1,681$	$4,921$
$D' = 1$ *	$3285,971$	$0,532$	$19824,780$	$3,209$	$6,033$

Tabela 1: Comparação no tempo total de execução, em segundos, utilizando configurações distintas para o parâmetro t, do programa para busca por LDs múltiplos nos dados do `SUCEST`. Foram realizados testes com $t = 5$ e $t = 60$. O parâmetro t define o tempo máximo, em segundos, de busca por uma clique de determinado tamanho. Maiores informações encontram-se no texto.

Quanto a SNPs, observamos dois parâmetros: maior clique possível (de acordo com seu grau e os graus dos vizinhos) e tamanho da clique encontrada. Obviamente, o primeiro parâmetro não varia. Já no segundo, percebemos um pequeno número de variações. O maior número concentrou-se nos casos relativos a métrica D'.

Ressaltamos aqui os números. Observe que foram comparados os resultados de 6178 *contigs*, com um total de 39608 SNPs, em quatro situações distintas. Não foram encontradas diferenças em relação às componentes conexas. Em apenas 507 testes, de um total de 158432, tivemos um incremento no tamanho da clique, ou seja, em apenas $0,32\%$ dos casos.

Chegamos a conclusão de que $t = 5$ segundos é suficiente para apresentar resultados satisfatórios.

3.3 Resultados

Doravante chamaremos as componentes conexas do grafo de "grupos de SNPs relacionados indiretamente", e as cliques de "grupos de SNPs relacionados diretamente".

As Figuras 3, 4, 5 e 6 nos apresentam gráficos comparativos das quatro situações estudadas.

No gráfico da Figura 3, para cada *contig* montamos o grafo e contamos as componentes conexas. Depois agrupamos os *contigs* pelo número de componentes conexas. O número de *contigs* é cumulativo. As médias obtidas foram $1,38 \pm 0,66$, $1,32 \pm 0,60$, $2,54 \pm 2,18$ e $2,38 \pm 1,95$ respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste grafo, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma maior tendência a criação de componentes conexas maiores, o que leva a um menor número de componentes conexas.

No gráfico da Figura 4, para cada *contig* montamos o grafo e calculamos a maior clique. Depois agrupamos os *contigs* pela tamanho da maior clique. O número de *contigs* é cumulativo. As médias obtidas foram $4,34 \pm 1,96$, $4,38 \pm 1,91$, $3,25 \pm 1,79$ e $3,34 \pm 1,88$ respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma maior tendência a formação de cliques maiores.

No gráfico da Figura 5, para cada *contig* montamos o grafo e contamos as componentes conexas com apenas um vértice. Depois agrupamos os *contigs* pelo número de componentes

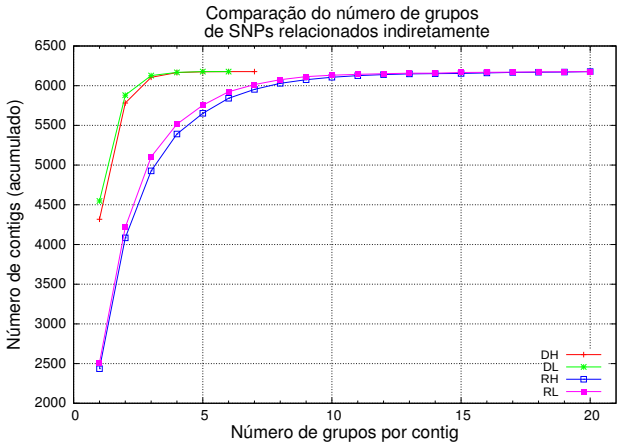

Figura 3: Gráfico comparando o número de grupos de SNPs relacionados indiretamente nos diversos casos estudados. No eixo X temos o número de componentes conexas (grupos de SNPs relacionados indiretamente) por *contig* e no eixo Y o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

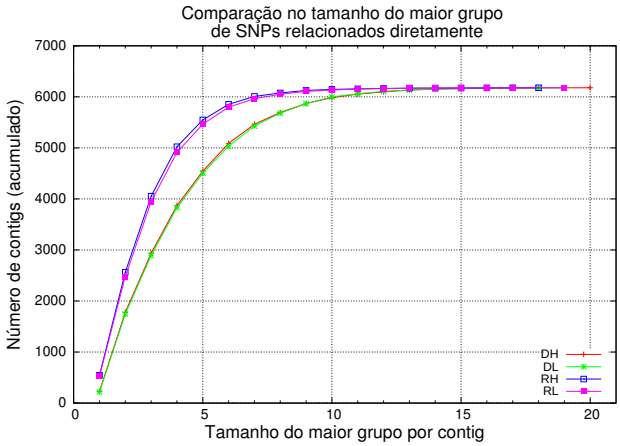

Figura 4: Gráfico comparando o tamanho do maior grupo de SNPs relacionados diretamente nos diversos casos estudados. No eixo X temos o tamanho da maior clique (grupo de SNPs relacionados diretamente) por *contig* e no eixo Y o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

unitárias. O número de *contigs* é cumulativo. As médias obtidas foram $0,26 \pm 0,58$, $0,23 \pm 0,55$, $1,18 \pm 1,67$ e $1,08 \pm 1,54$ respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma tendência menor ao isolamento de vértices.

No gráfico da Figura 6, para cada *contig* montamos o grafo e contamos os vértices que possuem clique máxima de tamanho dois. Depois agrupamos os *contigs* pelo número vértices com clique máxima de tamanho dois. O número de *contigs* é cumulativo. As médias obtidas foram $0,80 \pm 1,13$, $0,76 \pm 1,09$, $1,71 \pm 2,08$ e $1,64 \pm 2,02$ respectivamente para os casos $D' = 1$, $D' = 1*$, $r^2 \geq 1/3$ e $r^2 \geq 1/3*$. Neste gráfico, podemos observar que, utilizando $D' = 1$ como parâmetro para definir LD, há uma tendência menor a formação de cliques de tamanho dois.

Os gráficos apresentados nas Figuras 3, 4, 5 e 6 nos mostram que as métricas D' e r^2 , com limiares 1 (SNP completo) e 1/3 (SNP útil) respectivamente, possuem variações mas a tendência é a mesma.

Na Tabela 2 apresentamos um resumo de valores obtidos, para os quatro casos estudados, nos cálculos de LDs múltiplos nos dados da cana-de-açúcar. Valores para cinco parâmetros são listados, são eles: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP. Em primeiro lugar podemos observar que os valores considerando ou não bases de baixa qualidade são extremamente semelhantes. Em segundo lugar podemos validar tudo o que já havíamos observado nos gráficos. Utilizando a definição de LD completo há uma tendência de formação de menor número de componentes conexas, formação de componentes maiores, menor número de SNPs isolados e LDs simples.

Resultados obtidos nos casos estudados (SUCEST)

Parâmetro	$D' = 1$	$D' = 1*$	$r^2 \geq 1/3$	$r^2 \geq 1/3*$
Componentes Conexas (CCs)	1,4	1,3	2,5	2,4
Maior Clique (MC)	4,3	4,4	3,2	3,3
SNPs Isolados (C1)	0,3	0,2	1,2	1,1
LDs Simples (C2)	0,8	0,8	1,7	1,6
Média do Tamanho das Cliques (MTC)	4,0	4,1	2,8	2,9

Tabela 2: Comparação dos resultados dos cálculos de LDs múltiplos, nos dados da cana-de-açúcar no projeto SUCEST, utilizando a definição de LD completo ($D' = 1$) e de LD útil ($r^2 \geq 1/3$). O símbolo "$*$" indica que as bases de baixa qualidade foram consideradas nos cálculos. Na primeira coluna indicamos o parâmetro, na segunda apresentamos os resultados utilizando LD completo sem bases de baixa qualidade, na terceira utilizando LD completo com bases de baixa qualidade. As colunas 4 e 5 apresenta resultados semelhantes às colunas 2 e 3 só que agora utilizando a definição de LD útil. Os parâmetros listados são: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP.

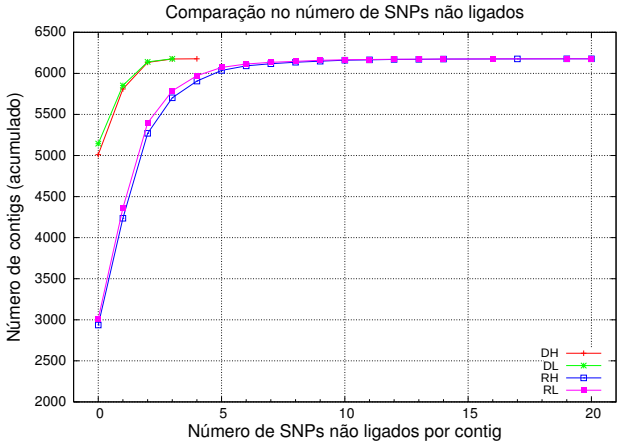

Figura 5: Gráfico comparando o número de SNPs não ligados nos diversos casos estudados. No eixo X temos o número de componentes conexas de tamanho um (SNPs não ligados) por *contig* e no eixo Y temos o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

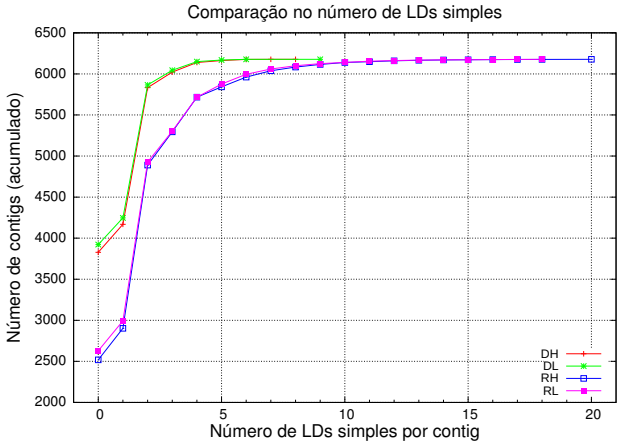

Figura 6: Gráfico comparando o número de LDs simples nos diversos casos estudados. No eixo X temos o número de vértices que possuem clique máxima de tamanho dois (LDs simples) por *contig* e no eixo Y temos o número de *contigs*. A curva DH representa o caso onde ocorre LD quando $D' = 1$ e bases de baixa qualidade são excluídas. A curva DL, quando $D' = 1$ e bases de baixa qualidade são consideradas. A curva RH representa o caso onde ocorre LD quando $r^2 \geq 1/3$ e bases de baixa qualidade são excluídas. A curva RL, quando $r^2 \geq 1/3$ e bases de baixa qualidade são consideradas.

Resumo dos dados coletados por gene				
	Antes da filtragem		**Após filtragem**	
	ESTs	**SNPs**	**ESTs**	**SNPs**
HLA-A	2535	2109	1475	280
HLA-B	2503	336	1740	144
HLA-DOB	42	77	42	41

Tabela 3: Número de ESTs e SNPs obtidos para cada gene selecionado da região MHC do cromossomo 6 humano, antes e depois do processo de filtragem. Os ESTs foram obtidos em formato `fasta` e os SNPs em formato `flat file`.

4 LDs múltiplos no genoma humano

Aplicamos também o algoritmo que descrevemos na Seção 2.1 em dados referentes a genes presentes no genoma humano. Selecionamos genes do cromossomo 6, de uma região conhecida como MHC, ou *Major Histocompatibility Complex*. Esta região de aproximadamente 4 Mbp é muito densa em genes, tendo mais de 120 genes, dos quais 40% codificam proteínas relacionadas a funções imunológicas. Mais de 20000 artigos foram escritos nos últimos 30 anos estudando a relação dos genes desta região com doenças [9, 18].

Através do site do NCBI [14] foram obtidos três genes, selecionados por se encontrarem em regiões com alta densidade de polimorfismos dentro do MHC: HLA-A (3324 bp, da base 30.018.310 à base 30.021.633 na montagem de referência), HLA-B (3287 bp, da base 31.429.628 à base 31.432.914 na montagem de referência) e HLA-DOB (4236 bp, da base 32.888.527 à base 32.892.762 na montagem de referência).

Para cada um dos genes, foi obtida a lista de ESTs e cDNAs utilizados para efetuar a montagem da seqüência do cromossomo, assim como a lista de SNPs de referência disponíveis na base de dados dbSNP [6] marcados como pertencentes aos genes escolhidos. Os ESTs e cDNAs foram filtrados de forma a se obter um conjunto de sequências sem bases indefinidas (símbolos N). A lista de SNPs também foi filtrada, eliminando os seguintes elementos:

- INDELS.
- SNPs cujas posições não são compatíveis com os limites dos genes respectivos.
- SNPs onde nenhum alelo corresponde ao alelo da seqüência de referência do gene.
- Posições redundantes. Em alguns casos, foram encontradas variações alélicas diferentes anotadas na mesma posição do genoma. Neste caso as variações foram agrupadas, de forma a que a lista final tivesse posições únicas. Por exemplo: as variações A/C e A/G anotadas na posição 10 são agrupadas, sendo considerada a variação A/C/G.

A Tabela 3 sumariza os dados obtidos antes e após a filtragem, para cada gene.
Geramos três conjuntos de dados, conforme descritos a seguir:

1. **NCBI Filtrado**: este primeiro conjunto de dados refere-se a um subconjunto do conjunto de SNPs marcados na base de referência (NCBI). Conforme apresentamos na Tabela 4, inicialmente tínhamos 280, 144 e 41 SNPs respectivamente para HLA-A, HLA-B e HLA-DOB. Ao aplicarmos filtros restaram 98, 28 e 10 SNPs. Tais filtros removeram SNPs tri ou tetra aléticos, além de remover posições que não pudemos mais considerar SNPs, depois dos filtros que aplicamos inicialmente, pois não há mais bases discordantes. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 5.

2. **Simples**: o segundo conjunto foi obtido observando as seções transversais dos *contigs* que continham variações alélicas. Foram descartados SNPs tri e tetra alélicos e só foram considerados SNPs quando a base de menor freqüência ocorreu, pelo menos, duas vezes e possui, pelo menos, 1% de freqüências dentre as bases da coluna. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 6.

3. **Intersecção**: o último dos conjuntos de dados é composto pela intersecção dos dois primeiros. Apresentamos os resultados dos cálculos para LDs múltiplos na Tabela 7.

Número de SNPs por conjunto de dados

Gene	NCBI	NCBI Filtrado	Simples	Intersecção
HLA-A	280	98	137	37
HLA-B	144	28	95	11
HLA-DOB	41	10	22	6

Tabela 4: Lista o número de SNPs em cada um dos conjuntos de dados do genoma humano onde foram calculados os LDs múltiplos. NCBI Filtrado, Simples e Intersecção são os conjuntos de dados. NCBI foi um conjunto inicial de onde foi extraído o conjunto NCBI Filtrado.

Comparação dos resultados (NCBI Filtrado)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	5	28	3	5	15,7	53	11	44	11	3,4
HLA-B	6	6	3	9	3,5	21	2	16	12	1,4
HLA-DOB	3	3	0	3	2,7	7	2	5	5	1,5

Tabela 5: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$). Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP.

Comparação dos resultados (Simples)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	2	18	0	5	9,7	60	10	48	38	3,0
HLA-B	3	15	1	14	9,0	45	5	36	23	2,3
HLA-DOB	1	10	0	1	8,5	3	8	2	5	4,8

Tabela 6: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$). Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP.

Comparação dos resultados (Intersecção)

	$D' = 1$					$r^2 \geq 1/3$				
	CCs	MC	C1	C2	MTC	CCs	MC	C1	C2	MTC
HLA-A	2	12	1	2	8,6	17	7	11	2	3,2
HLA-B	3	4	0	4	3,1	8	2	5	6	1,5
HLA-DOB	2	3	1	1	2,5	4	2	3	3	1,5

Tabela 7: Comparação dos resultados dos cálculos de LDs múltiplos, nos genes selecionados do cromossomo 6 humano, utilizando a definição de LD completo ($D' = 1$) e LD útil ($r^2 \geq 1/3$). Na primeira coluna indicamos o gene e nos dois blocos que seguem apresentamos os números. O primeiro bloco refere-se ao dados com a definição de LD completo e o segundo a LD útil. Em cada um dos blocos temos os seguintes valores: número de componentes conexas, tamanho da maior clique, número de SNPs isolados, número de LDs simples e média do tamanho das cliques associadas a cada SNP.

Assim como os resultados com a cana-de-açúcar pudemos observar uma tendência a formação de um menor número de componentes conexas, formação de componentes maiores e criação de um menor número de SNPs isolados e LDs simples quando utilizamos a definição de LD completo.

5 Conclusão

De acordo com os resultados aqui apresentados, chegamos a conclusão de que a métrica D' com limiar 1 (definição de LD completo) é a mais adequada para o cálculo de LDs múltiplos. Como pudemos observar, a métrica apresentou uma maior capacidade de agrupamento e menor tendência ao isolamento de SNPs ou formação de LDs simples.

Pudemos observar também que a métrica r^2 com limiar 1/3 (definição de LD útil) apresenta uma tendência semelhante a definição de LD completo. Sendo assim, caso o

desempenho (em relação a tempo) seja a prioridade, a definição de LD útil oferece bons resultados, conforme observamos nos testes com os genomas da cana-de-açúcar e humano. Observe que o algoritmo utilizando a definição de LD útil pode apresentar um desempenho (em termos de tempo) pior que o algoritmo utilizando a definição de LD completo para determinados conjuntos de dados. O desempenho do algoritmo está diretamente relacionado à estrutura do grafo.

Referências

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, C. R. Merril H. Xiao, A. Wu, B. Olde, and R. F. Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 21:1651–1656, 1991.

[2] Berkeley Drosophila Genome Project, July 2005. http://www.fruitfly.org/.

[3] D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics Supplement*, 33:228–237, March 2003.

[4] A. J. Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.

[5] D. N. Cooper, E. V. Ball, and M. Krawczak. The human gene mutation database. *Nucleic Acid Research*, 26:285–287, 1998.

[6] National Center for Biotecnology Information SNP Database, March 2006. http://www.ncbi.nlm.nih.gov/SNP.

[7] B. Devlin and N. A. Risch. A comparison of linkage disequilibrium measures for fine scaling mapping. *Genomics*, 29:311–322, 1995.

[8] W. G. Hill and B. S. Weir. Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of the Human Genetics*, 54:705–714, 1994.

[9] R. Horton, L. Wilming, V. Rand1, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. Conover Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck. Gene map of the extended human MHC. *Nature Reviews Genetics*, 5:889–899, December 2004.

[10] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.

[11] R. M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, pages 85–103, 1972.

[12] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, 1999.

[13] R. C. Lewontin. The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, 49:49–67, 1964.

[14] National Center for Biotecnology Information, March 2006. http://www.ncbi.nlm.nih.gov/.

[15] J. J. Pasternak. *Genética Molecular Humana - Mecanismo das Doenças Hereditárias*. Editora Manole, first edition, 2002. Título original em inglês: An introduction to human molecular genetics: mechanisms of inherited diseases.

[16] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239:487–491, 1988.

[17] J. C. Setubal and J. Meidanis. *Introduction to Computional Molecular Biology*. PWS Publishing Company, 1997.

[18] C. A. Stewart, R. Horton, R. J. N. Allcock, J. L. Ashurst, A. M. Atrazhev, P. Coggill, I. Dunham, S. Forbes, K. Halls, J. M.M. Howson, S. J. Humphray, S. Hunt, A. J. Mungall, K. Osoegawa, S. Palmer, A. N. Roberts, J. Rogers, S. Sims, Y. Wang, L. Wilming, J. F. Elliott, P. J. de Jong, S. Sawcer, J. A. Todd, J. Trowsdale, and Stephan Beck. Complete MHC Haplotype Sequencing for Common Disease Gene Mapping. *Genome Research*, 14:1176–1187, 2004.

[19] The Sugar Cane EST Genome Project, July 2006. http://sucest.lbi.ic.unicamp.br.

[20] A. L. Vettore, F. R. da Silva, E. L. Kemper, and P. Arruda. The libraries that made SUCEST. *Genetics and Molecular Biology*, 406:151–157, 2001.

[21] D. B. West. *Introduction to Graph Theory*. Prentice Hall, 1996.